CENG 3420
Computer Organization & Design

Lecture 13: Memory Organization-1

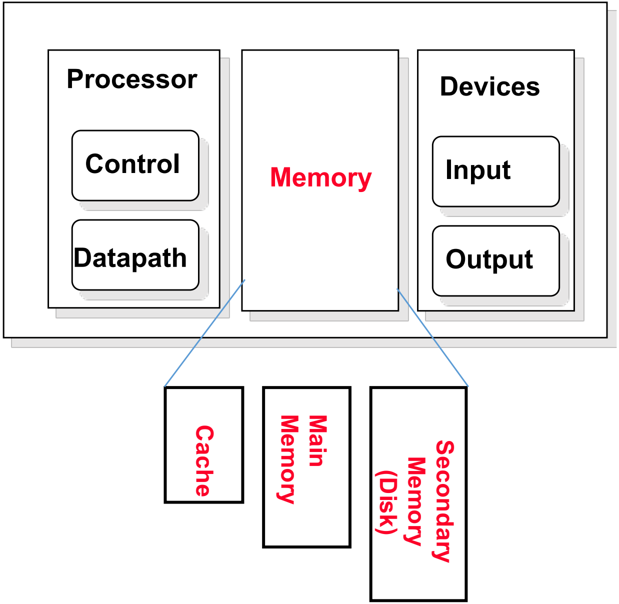
Bei Yu
CSE Department, CUHK
byu@cse.cuhk.edu.hk

(Textbook: Chapters 5.1–5.2 & A.8–A.9)

Spring 2022

Introduction

Review: Major Components of a Computer



Why We Need Memory?

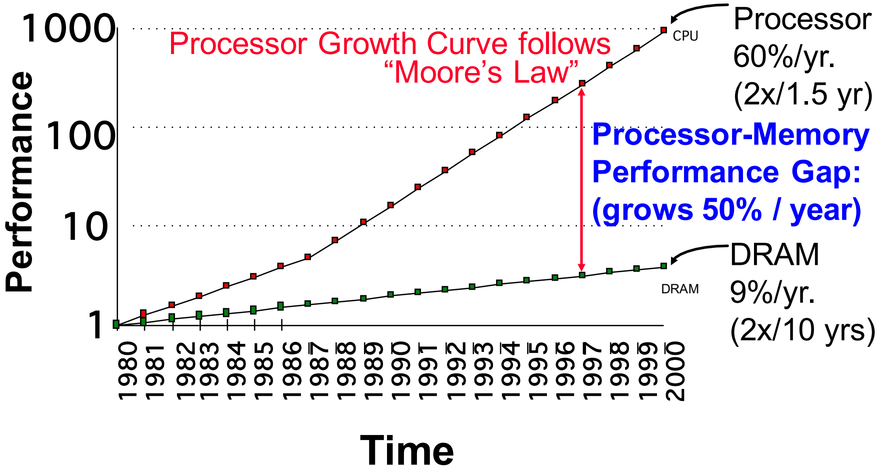
Combinational Circuit:

- Always gives the same output for a given set of inputs
- E.g., adders

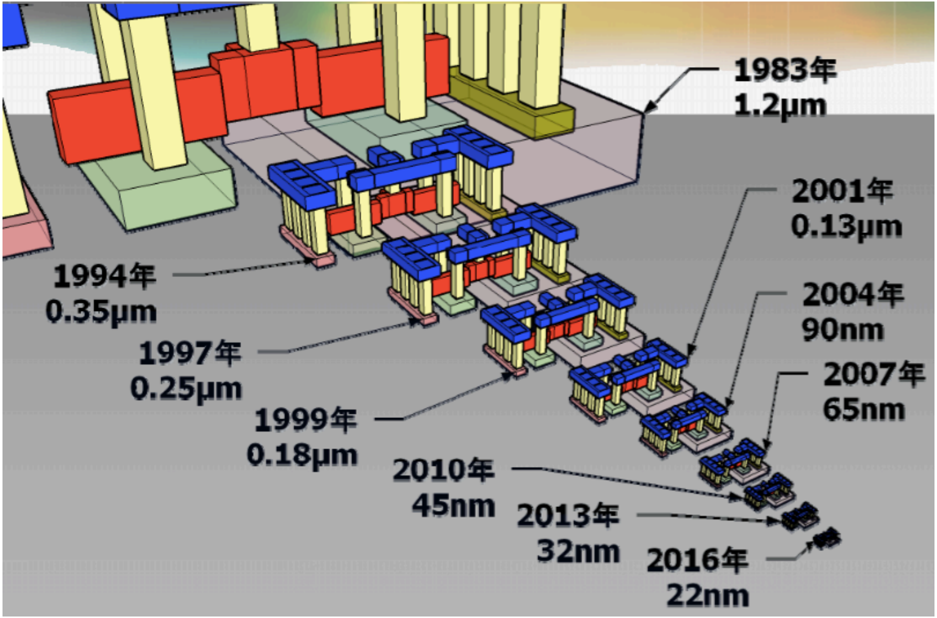
Sequential Circuit:

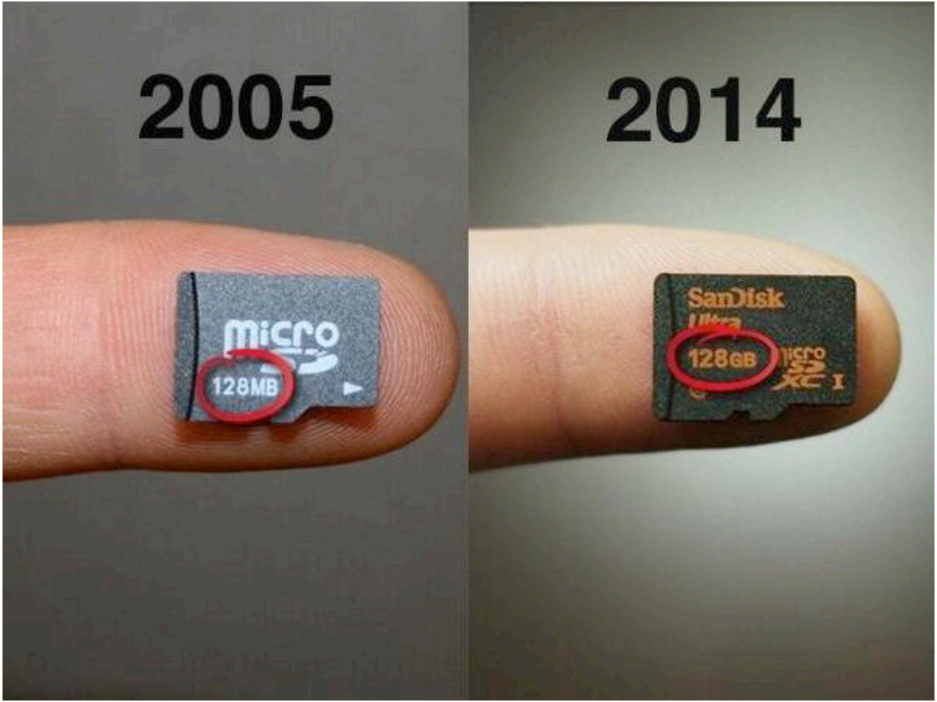
- Store information
- Output depends on stored information
- E.g., counter
- Need a storage element

Who Cares About the Memory Hierarchy?



Processor-DRAM Memory Performance Gap


1983年
1.2μm
2001年
0.13μm
1994年
0.35μm
2004年
90nm
1997年
0.25μm
2007年
65nm
1999年
0.18μm
2010年
45nm
2013年
32nm
2016年
22nm


2005
2014
micro
128MB
SanDisk
Ultra
128GB
micro SD XC I

Memory System Revisted

- Maximum size of memory is determined by addressing scheme

E.g.

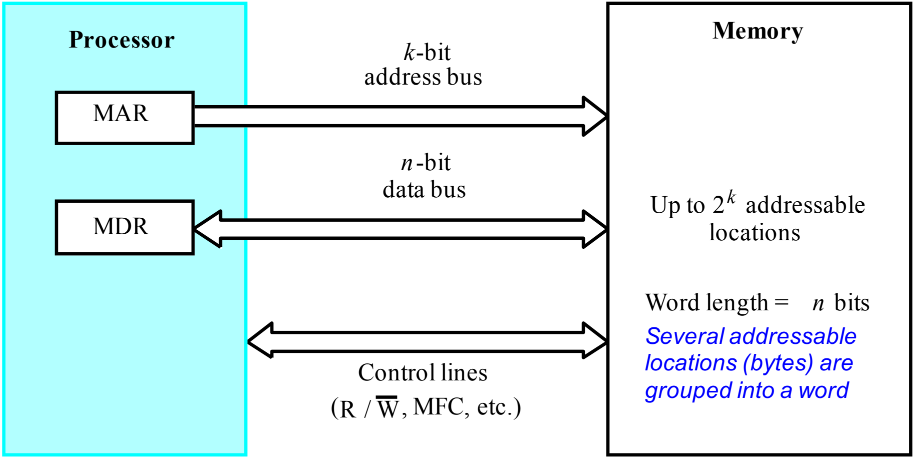
16-bit addresses can only address $2^{16} = 65536$ memory locations

- Most machines are byte-addressable
- each memory address location refers to a byte
- Most machines retrieve/store data in words
- Common abbreviations
  - $1\text{k} \approx 2^{10}$ (kilo)
  - $1\text{M} \approx 2^{20}$ (Mega)
  - $1\text{G} \approx 2^{30}$ (Giga)
  - $1\text{T} \approx 2^{40}$ (Tera)

Simplified View

Data transfer takes place through

- MAR: memory address register
- MDR: memory data register



Several addressable locations (bytes) are grouped into a word

Big Picture

Processor usually runs much faster than main memory:

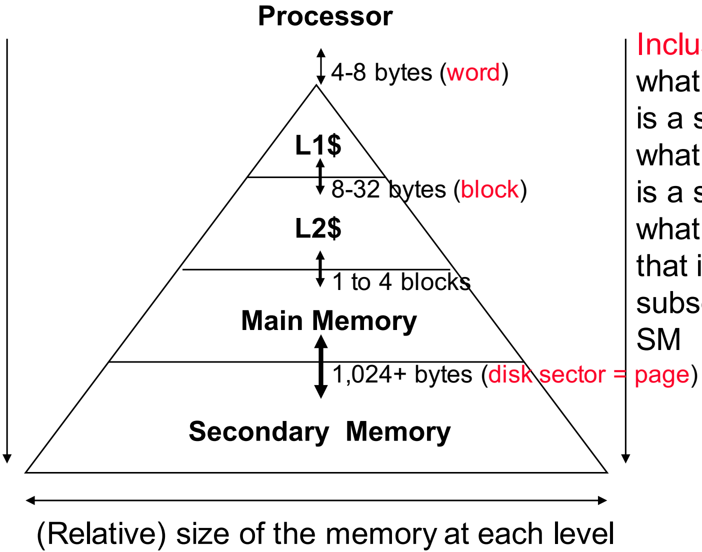
- Small memories are fast, large memories are slow.
- Use a cache memory to store data in the processor that is likely to be used.

Main memory is limited:

- Use virtual memory to increase the apparent size of physical memory by moving unused sections of memory to disk (automatically).
- A translation between virtual and physical addresses is done by a memory management unit (MMU)
- To be discussed in later lectures

Characteristics of the Memory Hierarchy

Increasing distance from the processor in access time



Inclusive– what is in L1$ is a subset of what is in L2$ is a subset of what is in MM that is a subset of is in SM

Memory Hierarchy: Why Does it Work?

Temporal Locality (locality in time)

If a memory location is referenced then it will tend to be referenced again soon

- Keep most recently accessed data items closer to the processor

Memory Hierarchy: Why Does it Work?

Temporal Locality (locality in time)

If a memory location is referenced then it will tend to be referenced again soon

- Keep most recently accessed data items closer to the processor

Spatial Locality (locality in space)

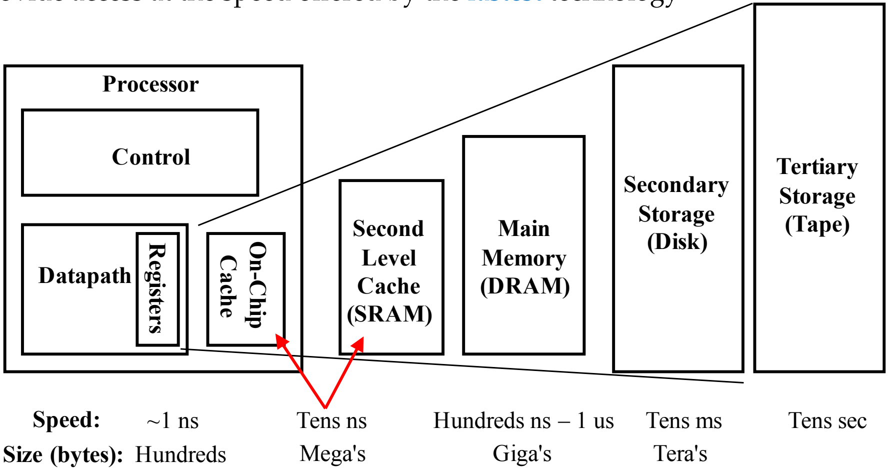
If a memory location is referenced, the locations with nearby addresses will tend to be referenced soon

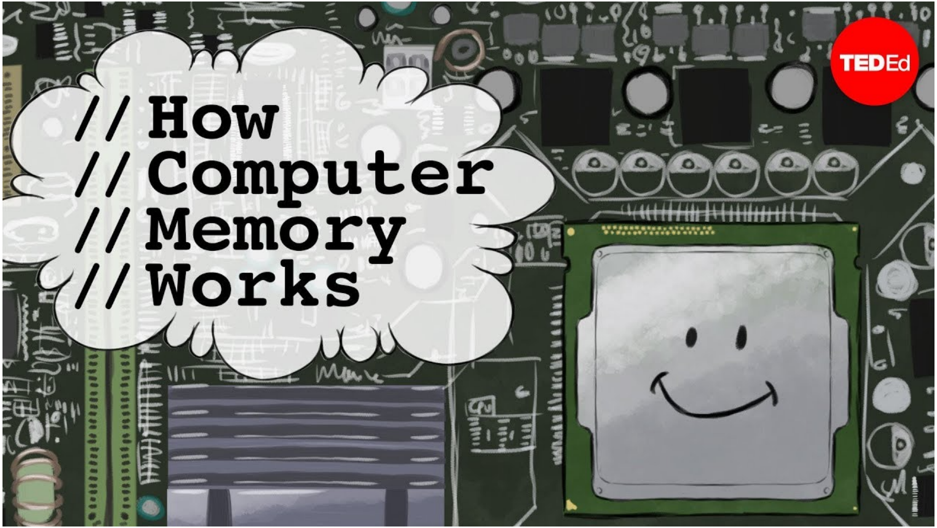
- Move blocks consisting of contiguous words closer to the processor

Memory Hierarchy

Taking advantage of the **principle of locality**:

- Present the user with as much memory as is available in the cheapest technology.
- Provide access at the speed offered by the fastest technology





https://youtu.be/p3q5zWCw8J4

Terminology

Random Access Memory (RAM)

Property: comparable access time for any memory locations

Block (or line)

the minimum unit of information that is present (or not) in a cache

Terminology

- Hit Rate: the fraction of memory accesses found in a level of the memory hierarchy
- Miss Rate: the fraction of memory accesses not found in a level of the memory hierarchy, i.e. 1 - (Hit Rate)

Hit Time

Time to access the block + Time to determine hit/miss

Miss Penalty

Time to replace a block in that level with the corresponding block from a lower level

Hit Time << Miss Penalty

Bandwidth v.s. Latency

Example

- Mary acts FAST but she’s always LATE.
- Peter is always PUNCTUAL but he is SLOW.

Bandwidth v.s. Latency

Example

- Mary acts FAST but she's always LATE.
- Peter is always PUNCTUAL but he is SLOW.

Bandwidth:

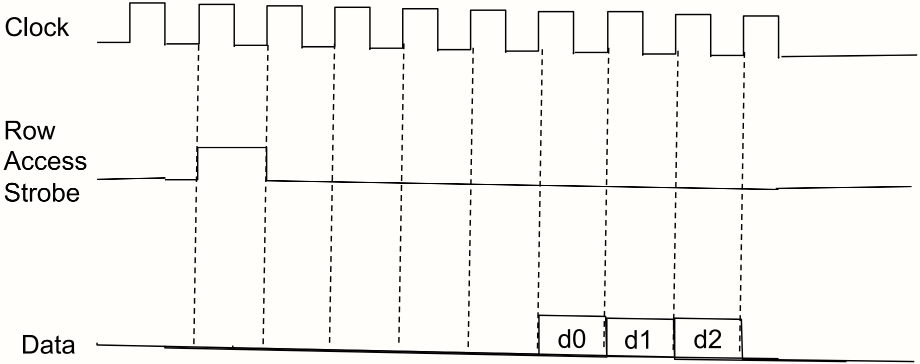
- talking about the "number of bits/bytes per second" when transferring a block of data steadily.

Latency:

- amount of time to transfer the first word of a block after issuing the access signal.
- Usually measure in "number of clock cycles" or in $ns/\mu s$.

Question:

Suppose the clock rate is 500 MHz. What is the latency and what is the bandwidth, assuming that each data is 64 bits?

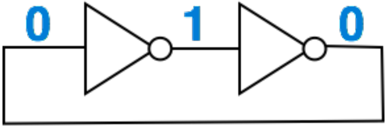


- 500 MHz = 2.0×10^{-9} second
- latency = 5 cycle = 10^{-8} second
- bandwidth = $\dfrac{8}{2 \times 10^{-9}} = 4 \times 10^{9}$ byte / second.

Information Storage

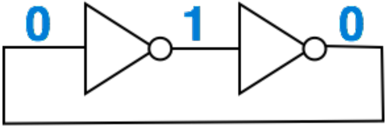
Storage based on Feedback

- What if we add feedback to a pair of inverters?

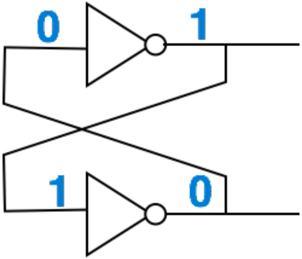


Storage based on Feedback

- What if we add feedback to a pair of inverters?

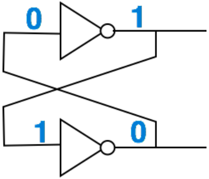


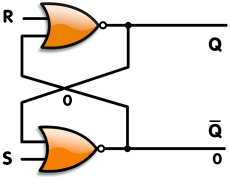
- Usually drawn as a ring of cross-coupled inverters
- Stable way to store one bit of information (w. power)



How to change the value stored?

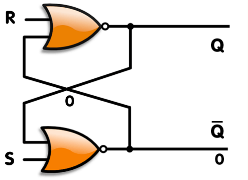
- Replace inverter with NOR gate
- **SR-Latch**

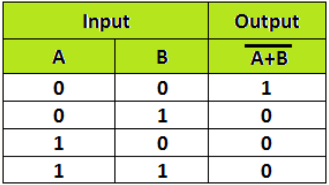




QUESTION:

What’s the Q value based on different R, S inputs?





Input		Output
A	B	$\overline{A+B}$
0	0	1
0	1	0
1	0	0
1	1	0

- R=S=1:
- S=0, R=1:
- S=1, R=0:
- R=S=0:

How to remember?

- S: set
- R: re-set

- R=S=1: not determined, not allowed
- S=0, R=1: set value to 0
- S=1, R=0:set value to 1
- R=S=0: latch holds current value